VideoPoet: A Large Language Model for Zero-Shot Video Generation

Year: 2023 | Citations: 381 | Authors: D. Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang

Abstract

We present VideoPoet, a language model capable of synthesizing high-quality video, with matching audio, from a large variety of conditioning signals. VideoPoet employs a decoder-only transformer architecture that processes multimodal inputs -- including images, videos, text, and audio. The training protocol follows that of Large Language Models (LLMs), consisting of two stages: pretraining and task-specific adaptation. During pretraining, VideoPoet incorporates a mixture of multimodal generative objectives within an autoregressive Transformer framework. The pretrained LLM serves as a foundation that can be adapted for a range of video generation tasks. We present empirical results demonstrating the model's state-of-the-art capabilities in zero-shot video generation, specifically highlighting VideoPoet's ability to generate high-fidelity motions. Project page: http://sites.research.google/videopoet/